REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
03-28-2012	Final Report	03/01/2009 - 11/30/2011

4. TITLE AND SUBTITLE
Optimization Techniques for Analysis of Biological and Social Networks

5a. CONTRACT NUMBER
FA9550-09-1-0154

5b. GRANT NUMBER
FA9550-09-1-0154

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Sergiy Butenko

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Dept. of Industrial and Systems Engineering, Texas A&M University, 4037 Emerging Technologies Build., College Station, TX 77843-3131, Phone: 979-458-2333, Fax: 979-458-4299, Email: butenko@tamu.edu

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Texas Engineering Experiment Station
1470 William D Fitch Pky
College Station, TX 77845-4645
(979) 458-7430

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2012-0948

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION A: APPROVED FOR PUBLIC RELEASE

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project focused on a multifaceted study of a class of cluster-detection problems arising in biological and social networks. This includes defining new cluster models and their alternative mathematical programming formulations, their theoretical analysis, the development of exact algorithms, and heuristics. Originally, clusters (complexes, modules, cohesive subgroups) in biological and social networks were described by cliques (complete subgraphs) or connected components. However, in many practical situations cliques appear to be overly restrictive, whereas connected components are insufficiently ``tight" clusters. This project considers a class of concepts describing clusters that ``relax" the definition of a clique and are tighter than connected components. Such problems are of great practical as well as theoretical interest and this project is the first attempt to approach the clique relaxation models in a systematic fashion under a unifying theoretical and algorithmic framework.

15. SUBJECT TERMS
Optimization, Complex Networks, Social Network Analysis, Computational Biology

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
				4	Sergiy Butenko
					19b. TELEPHONE NUMBER *(Include area code)* (979) 458-2333

Optimization Techniques for Analysis of Biological and Social Networks

PI: Sergiy Butenko
Dept. of Industrial and Systems Engineering, Texas A&M University, 4037 Emerging Technologies Build.
College Station, TX 77843-3131, Phone: 979-458-2333, Fax: 979-458-4299, Email: butenko@tamu.edu

Contract/Grant #: FA9550-09-1-0154

Final Project Report

03/01/2009 – 11/30/2011

Abstract

This project focused on a multifaceted study of a class of cluster-detection problems arising in biological and social networks. This includes defining new cluster models and their alternative mathematical programming formulations, their theoretical analysis, the development of exact algorithms, and heuristics. Originally, clusters (complexes, modules, cohesive subgroups) in biological and social networks were described by cliques (complete subgraphs) or connected components. However, in many practical situations cliques appear to be overly restrictive, whereas connected components are insufficiently "tight" clusters. This project considers a class of concepts describing clusters that "relax" the definition of a clique and are tighter than connected components. Such problems are of great practical as well as theoretical interest.

The developed approaches are based on representing the cluster-detection problems in biological and social networks in terms of discrete and continuous optimization models. In particular, one of the most promising directions in global optimization research deals with continuous (nonconvex) approaches to discrete optimization problems. The research in this project focused around the following three major thrusts:

1. Theoretical study of optimization models: (i) analyzing new optimality conditions for discrete optimization problems based on their continuous nonlinear formulations; (ii) using nonlinear formulations of discrete problems to deduct important computational complexity results for general and special cases of continuous optimization; (iii) obtaining lower and upper bounds based on nonlinear formulations that can be utilized in exact algorithms.
2. Algorithms development and analysis: (i) developing effective exact combinatorial algorithms for the problems of interest; (ii) designing new problem scale-reduction techniques and graph decomposition methods that can be used to solve special cases of large-scale instances arising in practical situations to optimality; (iii) developing and analyzing a new metaheuristic technique, *variable objective search*.
3. Experimentation and application: Implement the proposed algorithms, test and fine-tune the developed software using a variety of randomly generated problems and instances available in the public domain, and apply the software to real-life biological and social data.

The significance of clique and related concepts in the development of graph theory, theoretical computer science and optimization, including computational complexity theory and semidefinite programming, has been well documented. The models developed in this project provide systematic relaxations of the concept of clique and can potentially impact a number of important application areas.

1 Summary of Research Contributions

The research contributions resulting from the project are summarized in the following subsections. Namely, we list abstracts/brief summaries of the papers supported by this project that have been finalized and are published/accepted for publication or submitted/ready for submission.

In [1], we introduce and study the *maximum k-plex problem*, which arises in social network analysis, but can also be used in several other important application areas, including wireless networks, telecommunications, and graph-based data mining. We establish NP-completeness of the decision version of the problem on arbitrary graphs. An integer programming formulation is presented and basic polyhedral study of the problem is carried out. A branch-and-cut implementation is discussed and computational test results on the proposed benchmark instances and real-life *scale-free graphs* are also provided.

In [2], we study the maximum quasi-clique problem, which defines a cluster based on edge density. Given a simple undirected graph $G = (V, E)$ and a constant $\gamma \in (0, 1)$, a subset of vertices is called a γ-quasi-clique or, simply, a γ-clique if it induces a subgraph with the edge density of at least γ. The maximum γ-clique problem consists in finding a γ-clique of largest cardinality in the graph. Despite numerous practical applications, this problem has not been rigorously studied from mathematical perspective, and no exact solution methods have been proposed in the literature. This paper, for the first time, establishes some fundamental properties of the maximum γ-clique problem, including the NP-completeness of its decision version for any fixed γ satisfying $0 < \gamma < 1$, the weak heredity property, and analytical upper bounds on the size of a maximum γ-clique. Moreover, mathematical programming formulations of the problem are proposed and results of preliminary numerical experiments using a state-of-the-art optimization solver to find exact solutions are presented.

In [3], we introduce the variable objective search framework for combinatorial optimization. The method utilizes different objective functions used in alternative mathematical programming formulations of the same combinatorial optimization problem in an attempt to improve the solutions obtained using each of these formulations individually. The proposed technique is illustrated using alternative quadratic unconstrained binary formulations of the classical maximum independent set problem in graphs.

In [4] we deal with a diameter-based clique relaxation model. Given a simple undirected graph G, a k-club is a subset of vertices inducing a subgraph of diameter at most k. The maximum k-club problem (MkCP) is to find a k-club of maximum cardinality in G. These structures, originally introduced to model cohesive subgroups in social network analysis, are of interest in network-based data mining and clustering applications. The maximum k-club problem is NP-hard, moreover, determining whether a given k-club is maximal (by inclusion) is NP-hard as well. This paper first provides a sufficient condition for testing maximality of a given k-club. Then it proceeds to develop a variable neighborhood search (VNS) heuristic and an exact algorithm for MkCP that uses the VNS solution as a lower bound. Computational experiments with test instances available in the literature show that the proposed algorithms are very effective on sparse instances and outperform the existing methods on most dense graphs from the testbed.

Paper [5] analyzes the elementary clique-defining properties implicitly exploited in the available *clique relaxation* models and proposes a taxonomic framework that not only allows to classify the existing models in a systematic fashion, but also yields new clique relaxations of potential practical interest. Some basic structural properties of several of the considered models are identified that may facilitate the choice of methods for solving the corresponding optimization problems. In addition, bounds describing the cohesiveness properties of different clique relaxation structures are established, and practical implications of choosing one model over another are discussed.

In [6], we exploit the heredity property of two of the clique relaxation models, k-plex and s-defective clique, coupled with effective scale-reduction procedures, in order to develop effective exact algorithms for these models. The general node deletion problem can be stated as follows: Given a graph property Π, find the minimum number of nodes that need to be removed from the graph in order to obtain an induced subgraph satisfying property Π. In 1978, Yannakakis has shown that if the property Π is hereditary on induced subgraphs, nontrivial, and interesting, the resulting node-deletion problem is NP-hard. Such node-deletion problems have numerous practical applications, including social network analysis and graph-based data mining. This paper proposes an exact approach to solve the node-deletion problems of interest, which can be viewed as an extension of some of the most successful exact algorithms for the classical maximum clique problem. The excellent performance of the approach is illustrated through a comprehensive computational study for the maximum k-plex problem and the maximum s-defective clique problems.

In [7], we explore scale reduction techniques that use the knowledge of common neighbors to obtain the maximum clique on very large-scale real life networks (a million nodes). Analytically, the technique has been shown to be very effective on power-law random graphs. Experimental results on graphs from the SNAP database (Collaboration networks, P2P networks, Social networks, etc) show our procedure to be much more effective than a regular peeling approach, helping us obtain the maximum clique in all the test cases.

2 Survey Articles

Publication [8] is an encyclopedia article that introduces the closely related maximum clique, maximum independent set, graph coloring, and minimum clique partitioning problems. The survey includes some of the most important results concerning these problems, including their computational complexity, known bounds, mathematical programming formulations, and exact and heuristic algorithms to solve them. Finally, book chapter [9] describes the origins of clique relaxation concepts arising in social network analysis and provides a brief overview of their mathematical programming formulations, algorithms for solving the corresponding optimization problems, and selected real-life applications of the models of interest.

References

[1] B. Balasundaram, S. Butenko, and I. Hicks. Clique relaxations in social network analysis: the maximum k-plex problem. *Operations Research*, 59: 133–142, 2011.

[2] J. Pattillo, A. Veremyev, S. Butenko, V. Boginski. On the maximum quasi-clique problem. Conditionally accepted to *Discrete Applied Mathematics*, 2012.

[3] S. Butenko, O. Yezerska, and B. Balasundaram. Variable objective search. *Journal of Heuristics*, 2012. To appear. DOI 10.1007/s10732-011-9174-2.

[4] S. Shahinpour and S. Butenko. Algorithms for the maximum k-club problem in graphs. *Journal of Combinatorial Optimization*, 2012. To appear. DOI: 10.1007/s10878-012-9473-z.

[5] J. Pattillo, N. Youssef, and S. Butenko. Clique relaxation models in network analysis: taxonomy, cohesiveness, and optimization. Working paper.

[6] S. Trukhanov, B. Balasundaram, S. Butenko, and C. Balasubramaniam. Exact algorithms for hard node deletion problems in graphs. Working paper.

[7] A. Verma, A. Buchanan, and S. Butenko. Maximum clique problem on very large scale sparse networks. Working paper.

[8] J. Pattillo and S. Butenko. Clique, independent set, and graph coloring. In J. Cochran, editor, *Encyclopedia of Operations Research and Management Science*, Wiley, pages 3150–3163, 2011.

[9] J. Pattillo, N. Youssef, and S. Butenko. Clique Relaxation Models in Social Network Analysis. In M.T. Thai and P.M. Pardalos (editors), *Handbook of Optimization in Complex Networks*, Springer, pages 143–162, 2012.